SYNTHESIS OF SPEAKER FACIAL MOVEMENT TO MATCH SELECTED SPEECH SEQUENCES

K. C. Scott, D. S. Kagels, S. H. Watson, H. Rem, J. R. Wright, M. Lee, K.J. Hussey

Jet Propulsion Laboratory,
California Institute of Technology
under contract with
National Aeronautics and Space Administration

ABSTRACT- A system is described which allows for the synthesis of a video sequence of a realistic-appearing talking human head. A phonetic based approach is used to describe facial motion; image processing rather than physical modeling techniques are used to create the video frames.

INTRODUCTION

Computer synthesis of realistic-appearing talking human figures is an ongoing area of research. Traditionally, video sequences of this type are produced by artists using manual techniques. The Actors™ system is a figure synthesis system that is designed to produce animations of a human head speaking. Actors is being developed under the Automated Speech Visualization Task at the United States National Aeronautics and Space Administration's (NASA) Jet Propulsion Laboratory (JPL).

Using the Actors system, the facial movements of a speaker for selected speech sequences have been synthesized to demonstrate the resulting realism of and to test the Actors visible speech model. Some synthesized speech sequences include: "welcome", "the quick brown fox", "insert tab A into slot B", "1 am Ethel Merman", and "your mother was a hamster and your father smelt of elderberries".

Manual techniques for producing video sequences of a person speaking are labor intensive. Production of a video sequence of a person talking is highly dependent upon the capability of the animator to both render the images realistically and to produce realistic motion. The Actors system has been designed and developed to reduce the human labor required while increasing the realism of both the image quality and figure motion.

Development of the Actors system facility for computer synthesis of realistic video sequences of a talking human head requires solutions to the problems of simulation of figure motion and figure representation. A 'figure" in the Actors system is the head and face of the person speaking in the synthesized video sequence. JPL has developed the necessary computer graphics technology to synthesize a realistic representation of a person talking. Currently, JPL is developing a visible speech model that expresses the relation between spoken phonemes and face/mouth shape. This model will produce realistic talking motion in the synthesized video.

The figure is represented by the Actors speaker database which is a set of digital pictures of an actual person. 1 he various records in the database represent articulation of the face over the range of face shapes desired to be reproduced in the synthesized video sequence. Generally, a single image represents a single face shape. The articulated face shapes correspond to the production of phonemes during speech,

Figure motion is achieved in the synthesized video sequence by translating the desired output speech in phonemes to a list of face shapes and interpolating between them. The interpolation process achieves figure motion which is both continuous and smooth using a resampling process known, commonly, as "morphing". T he visible speech model is particularly important to the production of realistic motion. It identifies not only the face shape associated with a spoken phoneme, but also the transitions between face shapes and contextual dependencies.

I-he Automated Speech Visualization task commenced in 1992 with the investigation of techniques for producing realistic video sequences of a human head speaking. Development of the Actors system began in 1993. The Actors system will advance toward a hybrid approach combining the database with parametric models of the face shape and speech-related motion.

ACTORS SYSTEM

Approach

The approach taken in the Actors system to producing a video sequence of a person talking was to bound the problem space to that shown in Figure 1. The synthesis process requires as inputs a source video of the speaker (i.e. figure) to be represented in the output video, the desired speech of the output figure represented phonetically, and the audio track of the desired speech. The output product is a video sequence, or animation, of the subject in the source video speaking the desired speech.

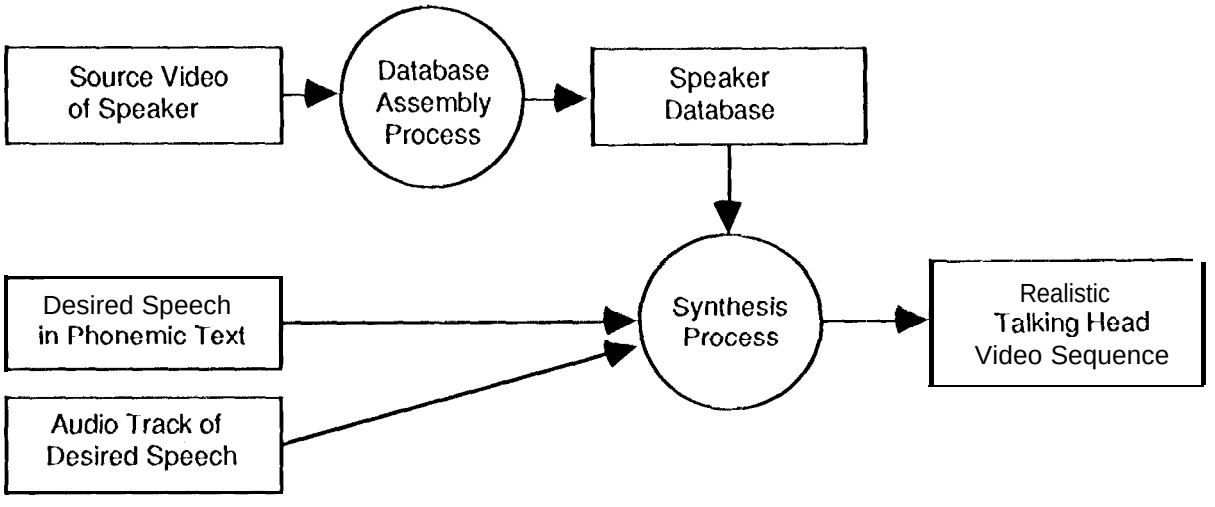

Figure 1. Problem domain of the Actors system

1 he source video of the speaker is a recorded video tape of the subject. T he subject is video taped speaking an arbitrary text that contains expression of the full list of desired face shapes, or database phonemes. The subject is recorded face-front speaking normally, recording both audio and video detail simultaneously. The desired speech is not represented, at the word or sentence level, in the context of the source video of the speaker. A sample text might contain a list of words, with each word or set of words containing a phoneme.

The audio track of the desired speech is the audio portion of the realistic audio-video output. It is a recording for which the video track will be synthesized in this process. The audio track may be synthesized by computer or recorded from a person, not necessarily the speaker in the source video.

The desired speech, in phonemic text, is a translation of the audio track into a sequence of phonemes. This sequence is a list of the phonemes actually spoken in the audio track and the time in the sound track when the phoneme is at maximum expression.

The realistic audio-video output can come in a variety of forms, but is generally a video tape with the synthesized video recorded in synchrony with the audio track.

Database Assembly Process

The Actors speaker database is a set of pictures that represent the range of face shapes through which the figure in the synthesized video sequence will move. The face shapes are digitized from the video track of the source video of the speaker. Using the audio track (and a phonemic transcript if desired), the specific video frames on the tape relating to each spoken phoneme are identified. These video frames are digitized and stored as pictures of the subject in the database on the computer.

Each picture represents a face shape and corresponds to a spoken phoneme. The face shape is digitized at the maximum expression or extreme of mouth motion/shape. Using the Actors T iepointer tool, an operator identifies a set of control points on the face which delineate the features (e.g. eyes, hair. mouth. etc.). The control points are used during the synthesis process to create smooth

transitions from one picture of the subject to another. Control points can be organized into groups to provide independent control of features. A display from the Actors Tiepointer tool showing the subject with control points and groups is shown in Figure 2.

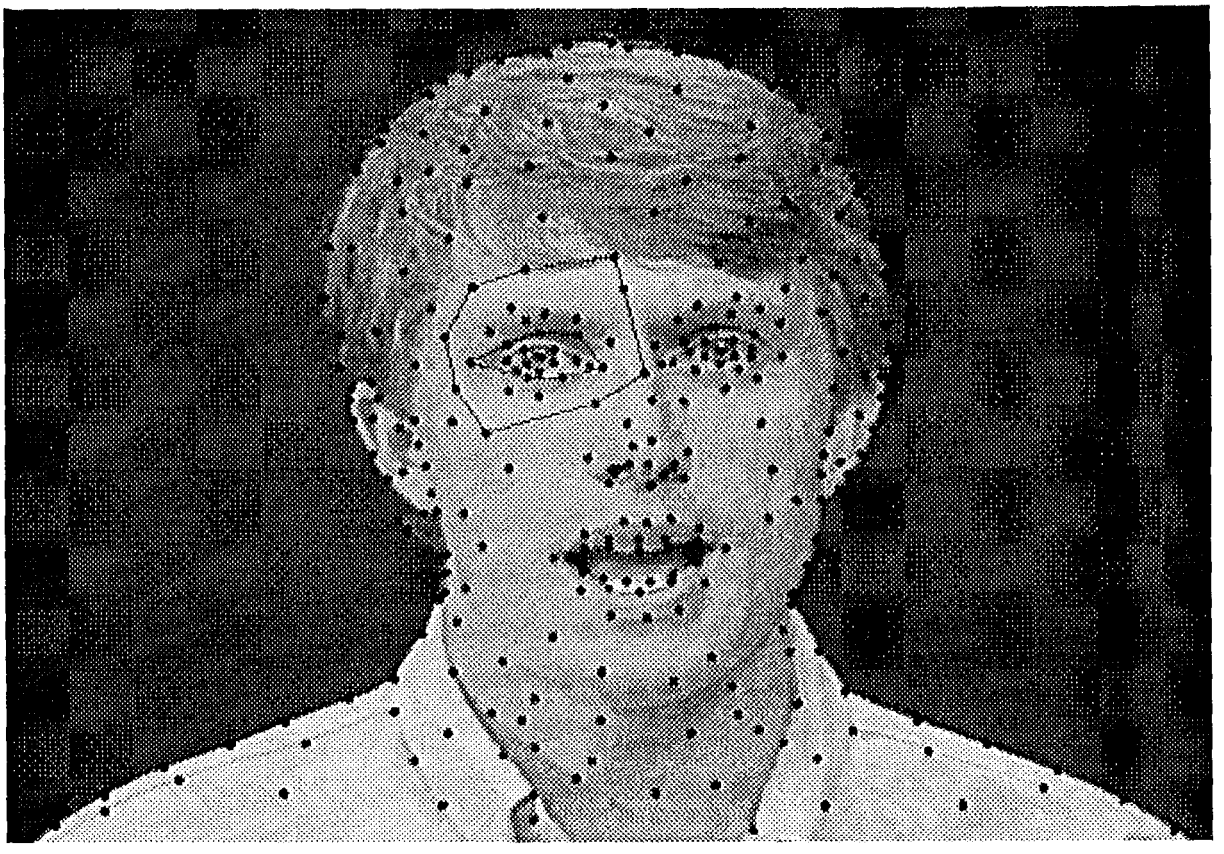

Figure 2. Actors Tiepointer tool showing a subject with control points and left eye groups

The number of records (face shapes) in the database depends upon the model relating face shape to phonemes, reduction due to similarities in face shapes, and other feature specific pictures (e.g. eyes closed, etc.). The visible speech model is described further in the section on figure motion.

Synthesis Process

Synthesis of the output video sequence is achieved through resampling of pictures in the speaker database. Resampling is more commonly known as morphing or rubber sheeting. The process allows for changing the shape of a picture without changing the content. The appearance of motion is affected by resampling and combining pictures in the database in linear combinations of face shapes over an interval of time.

A selected audio speech sequence (either synthesized *or* spoken) is the basis for synthesizing a matching video sequence. The speaker need not be the same as used for constructing the database. The audio sequence is analyzed to determine the spoken phoneme sequence and the relative timing of the enunciation of those phonemes.

l-he Actors Animator tool is used to specify precisely how to construct the output video sequence from the database of pictures, control points, and group definitions. The output video sequence is defined in terms of the audio sequence of spoken phonemes/times as a succession of key frames. Generally, a key frame corresponds to the expression of a phoneme. A key frame specification consists of the time of expression and linear combinations of both pictures and control points from the database. 1 he combination of pictures specifies the picture content of the synthesized frame, the combination of control points the shape of the picture content. Key frames can be added between phonemes to extend the interpolation from linear to piece-wise linear.

Changes in control point locations from key frame to key frame create the appearance of motion in the output video sequence. If the mouth is open in one key frame and closed in the next, the mouth will close linearly between the two key frames. Interpolation between face shapes with differing features is the basis for synthesis of motion. Interpolation between face shapes corresponding to phoneme expression gives the appearance of speech in the output video.

Grouping is a more advanced feature that allows for independent control of features of the face. Features that are specified as groups tend to be the head, eyes, and mouth. Entries into the speaker database for both control points and face shape can be specified from each group. Control of figure motion is provided through the Animator which allows the specification of parametric translation and rotation for each group.

FIGURE REPRESENTATION

The Actors system bases its representation of the figure on a database rather than a model, Model based approaches to this problem are being pursued in the computer graphics industry as a general solution to the problem, one that remains many years ahead in achieving realism. The database approach produces realistic appearing figures and motion.

1 he speaker database is the fundamental representation of the figure to be animated in the output video. The database is a set of pictures of the subject's head/face, with each picture having been digitized from the source video of the speaker. A limitation of this database driven approach is that only what is stored in the database can be expressed in the synthesized video sequence. The Actors synthesis process aflows for complex combinations of database records to be used in the production of an output picture, thereby increasing the possible output set to linear combinations of database elements.

T he figure is represented in the Actors speaker database as a set of digital pictures of an actual person. Each picture is a record in the database. The various records in the database represent articulation of the face over the range of face shapes desired to be reproduced in the synthesized video sequence. For speech related articulation each record corresponds to the production of a phoneme. Other records may relate to other facial characteristics such as eyefid motion (open and closed), eyeball look direction (up, down, left, and right), and emotion.

Figure representation is based on the visible speech model which relates the set of speech related records of face shape in the speaker database to the production of a spoken phoneme. The input to the model is the sequence of spoken phonemes, the output is a sequence of database records or combination of records that reproduce the correct face shape during phoneme utterance. The initial visible speech model used in the Actors system expressed this relationship as one-to-one, i.e. each spoken phoneme was represented by one unique face shape in the database. The phonemic coding scheme used 50 phonemes.

Based on the initial visible speech model, a full speaker database was produced and a set of animations were synthesized to demonstrate the resulting level of realism. An analysis of the results and examination of the database revealed that diphthongs are not adequately represented by a single face shape and that the database contains redundancies in face shape.

The production of a diphthong acoustically is a glide between two sounds. The start and end sounds are approximately that of two vowels, as a plot of F1 versus F2 formants clearly shows . Thus, visually, the shape of the face must also be represented as a glide between two face shapes. The visible speech model was extended to include representation of diphthongs as a glide between two face shapes, represented by the records in the speaker database corresponding to the production of the refevant two vowels, Sample video sequences were produced to test this hypothesis; the result was more realistic expression of the face shape to accompany the sound of a diphthong.

1 he speaker database of face shapes contained obvious redundancies. Two approaches to reducing the redundancies have been considered. First, eliminate redundancies based on characteristics of productions such as voiced/unvoiced pairs and location in vocal tract. Second, categorize the face shapes and eliminate commonality. Sample video sequences were synthesized based on substitution of voice/unvoiced pairs with no appreciable visual difference. Reduction based on categorized face shapes has not yet been tested.

FIGURE MOTION

Speech related facial motion in the Actors system is based on interpolation of face shapes, Face shapes are stored as a set of control points for each picture in the database. The control points identify the location of facial features for each face shape. Of particular importance to speech are the facial features that **vary in the production of speech. The main feature is the mouth, which includes** the lips, teeth, tongue, jaw, and cheeks.

The speaker database contains the set of face shapes over which the face must range to visually simulate speech. The major component of face shapes in the database are pictures of the subject speaking a full set of phonemes. The visual appearance of speech is produced by displaying in order and at the appropriate rate a sequence of face shapes based on a phonemic translation of the desired speech. To smooth the motion of the figure, the interval between face shapes is filled with frames synthesized by morphing from the face shape at the beginning of the interval to that at the end.

For example, if the word to be spoken is **"Poe",** translated as /p/ occurring at time A and /o/at time B, in the range of time between A and B the mouth will linearly transition from the pursed lip shape of the bilabial /p/ to the lip-rounded shape of the /o/. The /p/ picture is displayed at time A, the frames between A and B are synthesized from a linear combination of A and B by morphing, and the /o/ picture is displayed at time B. A display from the Actors Animator tool with a more lengthy example is shown in Figure 3.

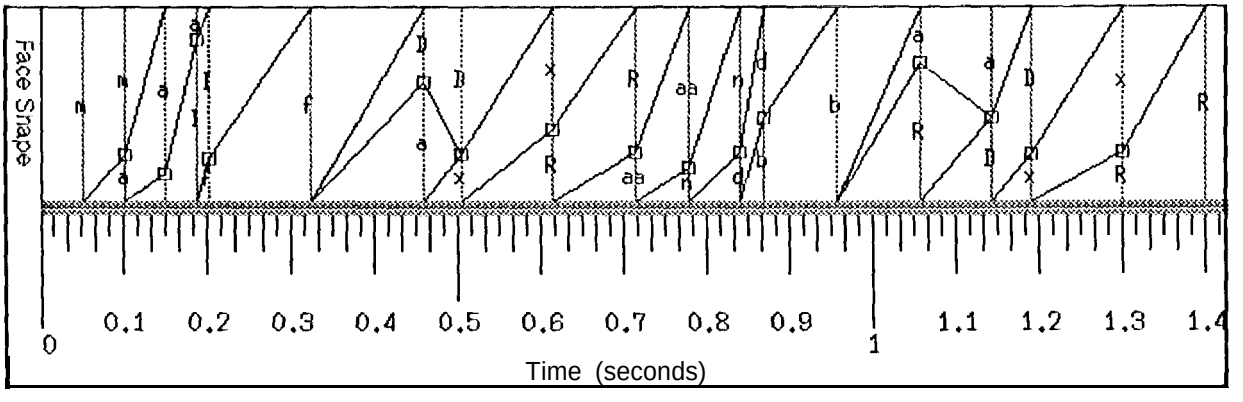

Figure 3. Actors Animator tool showing the face shape sequence
for the phrase "my father and brother"

The initial visible speech shape model fully expressed the face shape of each phoneme and linearly interpolated between phonemes. At maximum acoustic expression of the phoneme, the relevant face shape in the speaker database fully controlled the face shape in the synthesized output video. Test sequences showed that full visual expression of all phonemes has an unnatural appearance. Visually this resulted in unnaturally fast, jerky and extreme mouth motion.

The visible speech model was modified to base extent of visual expression on the location of sound production in the vocal tract. Generally, a sequence of phonemes is established that control the shape of the face. The controlling phonemes are produced mainly by the lips and teeth. Phonemes produced behind the teeth in the mouth cavity affect the shape of the face without controlling it. Phonemes produced, behind the vellum have no control or affect on face shape. The affect on face shape is accomplished in the Animator tool by establishing key frames that are a linear combination of face shapes, the majority percentage from the controlling phoneme and the minority percentage from the affecting phoneme.

Sample video sequences were produced, the result was more natural looking mouth motion. The conclusion drawn from this investigation is that certain phonemes do not have an associated face shape (i.e. face shape is irrelevant to the production of the sound) while others may have influence on face shape without controlling il. Also that visual expression of certain phonemes is sensitive to the context of the preceding and succeeding phonemes.

CONCLUSION

In conclusion, JPL has been successful in synthesizing the facial motion of a native American-English speaker for a small set of arbitrary speech segments. Much of the early work focused on successive graphical problems, including: registration of head position and head rotation to eliminate head jerks, stabilization of shoulders to eliminate registration-induced shoulder bobbing, removal of an induced, rubber neck motion, disturbing eye artifacts due to differences in recording, and induced motion of the background.

The Actors database approach to figure representation and motion produces highly realistic results, but has strict limitations on the range of reproducible facial expressions. Actors can be extended to a hybrid system which incorporates parametric models, retaining the realism of the current system while increasing the expressible range beyond that of the database.

1 **he** Actors visible speech model for relating speech to face shape produces realistic appearance and motion. Based on these results, a more formal model relating spoken phonemes to face shape is being developed in a joint research effort with the University of California at Los Angeles.

ACKNOWLEDGMENT

T his paper represents one phase of research performed at the Jet Propulsion Laboratory, California Institute of Technology through an agreement with the National Aeronautics and Space Administration.

This work has been supported with patience, tolerance, and ready availability by Steve Groom, the chief subject (see Figure 2).

REFERENCES

Rabiner, L.R. & Schafer, R.W. (1978) *Digital* Processing *of Speech Signals,* (Prentice-Hall)

Parke, F.I. (1974) *A Parametric Mode/fort luman faces,* **(University of Utah [dissertation])**

(as identified in the DecTalk documentation based on phonetic transcription work by Terry Sejnowski **[where at? or identify reference to paper In Dectalk doc]** and **Charles Rosenberg of Princeton)**

[reference to face shape paper ?]